Omega 39 (2011) 550–557

Contents lists available at ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

Practical scheduling for call center operations

Dennis C. Dietz

Qwest Communications International Inc., 1855 South Flatiron Court, Boulder, CO 80301, USA

ARTICLE INFO

Article history:
Received 29 July 2009
Accepted 11 December 2010
Processed by B. Lev
Available online 17 December 2010

Keywords:
Queueing
Quadratic programming
Scheduling
Operations management

ABSTRACT

A practical spreadsheet-based scheduling method is developed to determine the optimal allocation of service agents to candidate tour types and start times in an inbound call center. A stationary Markovian queueing model with customer abandonment is employed to determine required staffing levels for a sequence of time intervals with varying call volumes, handling times, and relative agent availabilities. These staffing requirements populate a quadratic programming model for determining the distribution of agent tours that will maximize the fraction of offered calls beginning service within a target response time, subject to side constraints on tour type quantities. The optimal distribution is scaled to reflect the total number of scheduled agents, and a near-optimal integer solution is derived using rounding thresholds found by successive one-dimensional searches. This novel approach has been successfully implemented in large service centers at Qwest Communications and could easily be adapted to other operational environments.

1. Introduction

Many commercial enterprises and public agencies operate centralized call centers to provide effective and responsive service for patrons. Mandelbaum [1] estimates that there are as many as 200,000 separate call centers operating in the United States, employing up to 4% of the national workforce (more than the entire agricultural sector). About 70% of the operating cost of a typical call center is attributable to personnel expense, so the economic efficiency of the operation is determined primarily by the quality of the employee scheduling process. For inbound call centers, the scheduling problem is normally characterized by a highly variable demand pattern and a requirement to assign service agents to "tours" that are constrained by labor rules. The fundamental problem is to schedule tours such that resulting time-varying staff quantities maximize the service level, or achieve a target service level at minimum cost.

Efficient management of a modern call center involves decision making (and supporting modeling and analysis) on three primary time horizons: annual planning, monthly (or quarterly) scheduling, and daily execution. Annual planning deals with strategic concerns such as forecasting long-term call volume trends and associated personnel requirements, managing an employee replacement pipeline, planning for volume seasonality, and conducting an annual vacation bid. Daily execution encompasses tactical matters such as consideration of schedule change requests, monitoring of schedule compliance and center performance metrics, and responding to unpredicted fluctuations in call volume by offering discretionary time-off or overtime to appropriate agents. This article focuses on monthly scheduling, which involves confirming forecast volume and total staff quantities, adjusting for nonproductive activity requirements (estimating agent "availability"), creating a schedule, and then populating the schedule with particular employees based on seniority and preferences. We are specifically concerned with the technical task of creating an optimal schedule, which is derived as an optimal quantification of tours by type and start time.

The importance of the call center scheduling problem is indicated by a large and growing body of relevant literature. Gans et al. [2] present a cogent overview, and Mandelbaum [3] provides a comprehensive bibliography. Reported application areas include retail sales [4], transportation [5], public services [6], and the telecommunications industry [7,8]. Solution approaches have incorporated diverse management science methods such as mathematical programming [9,10], analytical queueing models [11], simulation [12], dynamic programming [13], genetic algorithms [14], and other heuristic procedures [15]. Brigandi et al. [16] document deployment of a call center modeling system that delivered $750 million in increased annual profits for a diverse set of client enterprises. The system relied on simulation as the primary modeling tool, but employed queueing models to calculate staffing requirements and a network flow approach to determine workforce schedules. In this article, we apply queueing theory, quadratic programming, and a one-dimensional search algorithm to derive and evaluate optimal schedules, all within a practical spreadsheet implementation.

2. Determining staffing requirements

A forecast weekly demand profile for a typical call center can be accurately constructed from historical data. Fig. 1a displays an

E-mail address: dennis.dietz@qwest.com

0305-0483/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.omega.2010.12.001

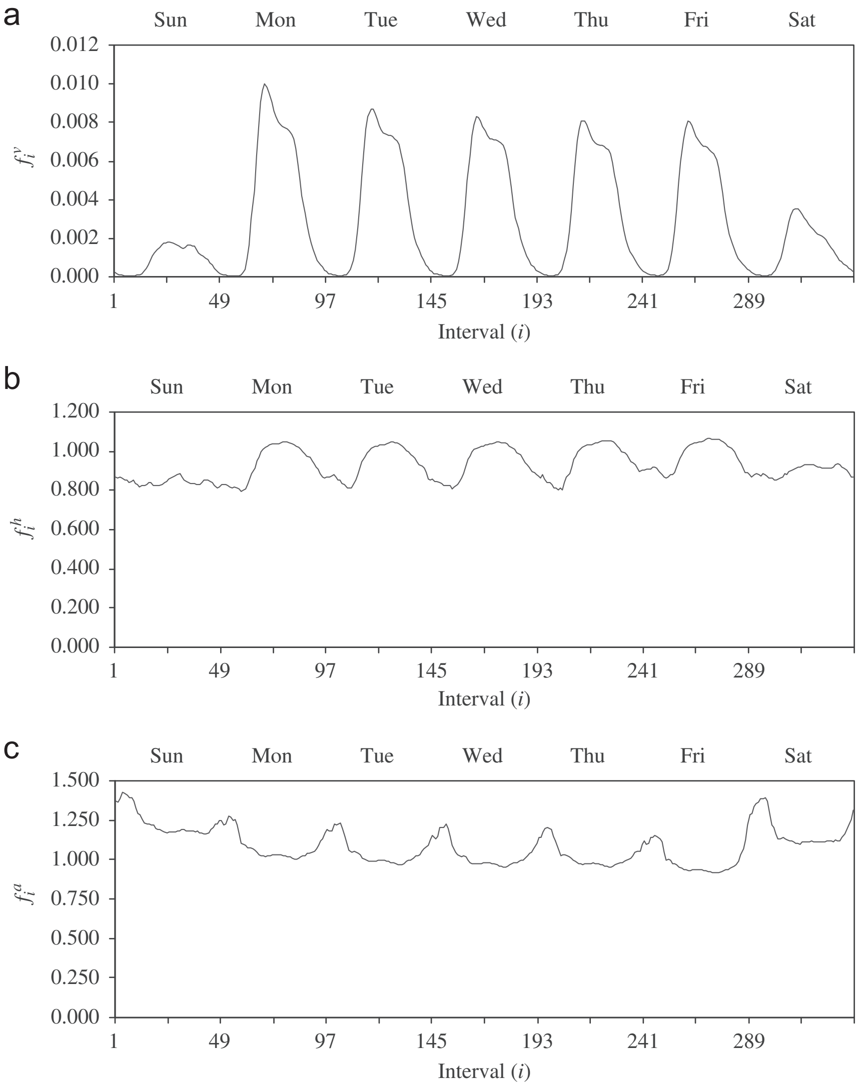


Fig. 1. Typical parameter profiles. (a) Offered call volume, (b) average handling time and (c) staff availability.

expected distribution of offered repair calls for a typical week and product at Qwest Communications. For any future week, the expected call volume ν_i for each 30-min operating interval $i \in I \subseteq \{1,\ldots,336\}$ is determined as the product of the associated profile value f_i^v (where $\sum_{i\in I} f_i^v = 1$) and a forecast weekly volume V. For the particular product depicted, the weekly volume varies seasonally by about 30% from its low value in December to its peak value in August. The distribution of call volume among intervals within the week, however, is demonstrably invariant throughout the year. Variability in realized call volume within an interval can be treated as random, so the customer arrival process can be modeled as a nonstationary Poisson process with an expected number of arrivals ν_i. A similar approach is pursued to capture interval-dependent variation in handling time. Fig. 1b displays handling time profile values f_i^h which are aggregated from annual interval data and scaled such that $\sum_{i\in I} f_i^v f_i^h = 1$. The profile indicates the presence of recurring patterns including "shift fatigue" (longer service time during high volume intervals), which is commonly discerned [11]. Letting H be a specified average handling time for a given future week, average handling time for each interval i can be computed as $h_i = f_i^h H$ (the scaling of f_i^h ensures that $\sum_{i\in I} f_i^v h_i = H$). Finally, Fig. 1c displays a staff availability profile, which captures interval-dependent variability in the average fraction of time a scheduled agent is actually available to handle calls after accounting for nonproductive activities such as absences, breaks, meetings, training, and other administrative functions. The availability factor for interval i is computed as $a_i = f_i^a A$, where the profile values f_i^a are similarly derived from annual interval data and A is the average availability estimate for the week (A must be a number between 0 and $1/\max_{i\in I}\{f_i^a\}$, so that $0 \le a_i \le 1, i \in I$). Since an efficient schedule will correlate interval staffing levels with corresponding work volumes, the values f_i^a are scaled to ensure that $\sum_{i\in I} f_i^v f_i^h f_i^a = 1$ (so that $\sum_{i\in I} f_i^v f_i^h a_i = A$). By decoupling H and A from their associated profiles, we can conveniently model trends and seasonalities in these factors which do not appreciably affect their relative magnitudes across intervals. We note that all three profiles must be periodically and simultaneously updated due to interaction between call volume, handling time, available staff, and implemented schedules.

When the scheduling objective is to minimize total cost (surrogated by staff size), an optimal schedule must ensure that sufficient agents are assigned on each interval to satisfy a composite service level requirement for each week of the relevant scheduling period. Alternatively, when the staff size is specified, the service level requirement can be iteratively adjusted until the predetermined number of agents is employed in the optimal solution. For scheduling purposes, we narrowly define service level as the probability that a random customer will not wait more than a

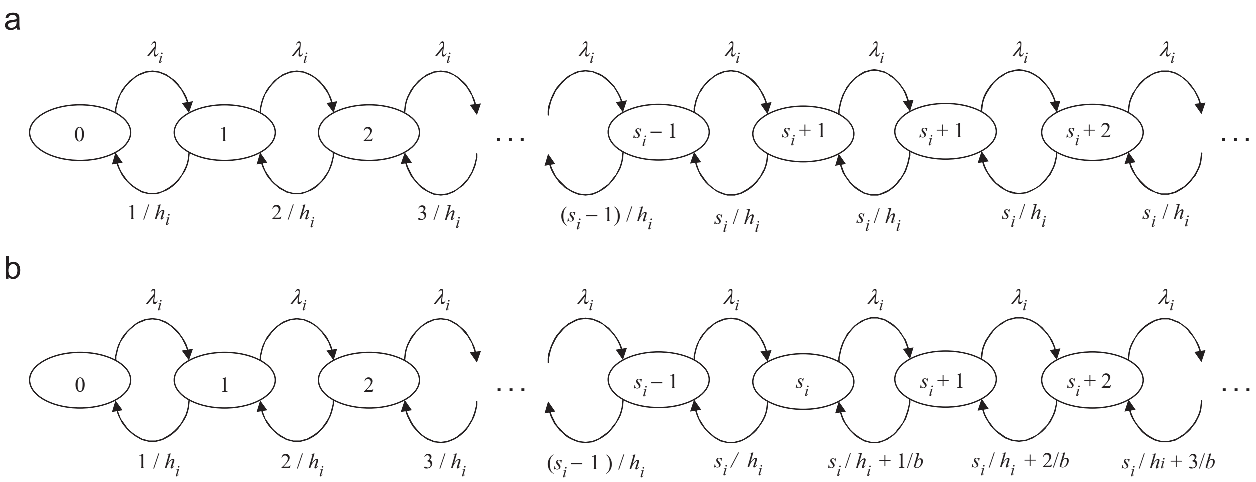


Fig. 2. Transition diagrams for queueing models. (a) Conventional model and (b) improved model (includes customer abandonment).

specified time for agent contact. The probabilistic nature of the call arrival and handling process suggests the application of an analytical queueing model in predicting performance measures. The standard approach when service times are short and the service level requirement is high is to assume that each interval quickly approximates steady-state conditions, and that the intervals can be analyzed independently (the SIPP, or Stationary Independent Period-by-Period, assumption). Green et al. [17,18] describe conditions under which the SIPP assumption is reasonable and suggest analytical remedies for variant situations. Under SIPP, an interval with s_i available agents can be approximately modeled by assuming an exponential distribution of handling times with mean of h_i seconds, and a Poisson call arrival rate of v_i per interval or $\lambda_i = v_i/1800$ per second. A transition diagram for the resulting Markovian birth–death process, commonly referred to as an "Erlang C" model, is displayed in Fig. 2a. Letting $P_i(n)$ represent the stationary probability that n customers are in the system, local balance equations for the process can be written as

$$P_i(n+1) = P_i(n)\lambda_i h_i/(n+1), \quad 0 \le n < s_i \tag{1}$$

$$P_i(n+1) = P_i(n)\lambda_i h_i/s_i, \quad s_i \le n < \infty \tag{2}$$

$$\begin{aligned} 1 = \sum_{n=0}^{\infty} P_i(n) &= \sum_{n=0}^{s_i-1} P_i(n) + P_i(s_i)\sum_{n=0}^{\infty}(\lambda_i h_i/s_i)^n \\ &= \sum_{n=0}^{s_i-1} P_i(n) + \frac{P_i(s_i)}{1-\lambda_i h_i/s_i} \end{aligned} \tag{3}$$

for $\lambda_i h_i/s_i < 1$. Since $\sum_{n=s_i}^{\infty} P_i(n)$ can be replaced by a scaled geometric series, closed-form expressions can be derived for all standard performance measures (see Cooper [19, pp. 90–102]). For example, letting W_i represent customer waiting time and t be the target response time (e.g., 20 s), service level is computed as

$$\begin{aligned} P\{W_i \le t\} = 1 - &\left\{\frac{(\lambda_i h_i)^{s_i}\exp\{(\lambda_i - s_i/h_i)t\}}{(s_i-1)!(s_i-\lambda_i h_i)}\right\} \\ &\times \left\{\sum_{n=0}^{s_i-1}\frac{(\lambda_i h_i)^n}{n!} + \frac{(\lambda_i h_i)^{s_i}}{(s_i-1)!(s_i-\lambda_i h_i)}\right\}^{-1}. \end{aligned} \tag{4}$$

To determine an interval staffing requirement, we can initialize the staffing level at $s_i = \lceil \lambda_i h_i \rceil$ and then increment s_i until $P\{W_i \le t\}$ exceeds a specified service level.

In modeling Qwest repair call handling centers, conservatism of the conventional queueing model has been verified through simulation-based performance analysis [20]. A significant weakness of this approach is that customers are assumed to possess infinite patience and, hence, can depart the queue only by entering service. Consequently, when $\lambda_i > s_i/h_i$ for any interval, an infinite queue is predicted. Such intervals prohibit computation of an average queue length or waiting time for the entire week. Furthermore, in heavy traffic, even a small fraction of abandoning customers can dramatically affect system performance. For these reasons, explicit modeling of customer abandonment offers an important improvement to the conventional model [2,21].

Analytical methods for incorporating customer abandonment in queueing systems were first considered by Palm [22], and are described by several authors including Riordan [23], Garnett et al. [24], Stolletz [25], Feldman et al. [26], and Whitt [27]. Some of these contributors consider general probability distributions for handling time and customer patience, but Brown et al. [21] demonstrate that corresponding exponential models are often quite robust. Fig. 2b displays a transition diagram for the revised birth–death process, which is referred to as an "Erlang A" model in some recent call center literature [2,18,28]. The model incorporates an exponentially distributed customer patience with mean b, so the associated balance equations are

$$P_i(n+1) = P_i(n)\lambda_i h_i/(n+1), \quad 0 \le n < s_i \tag{5}$$

$$P_i(n+1) = P_i(n)\lambda_i/\{s_i/h_i + (n+1-s_i)/b\}, \quad s_i \le n < \infty \tag{6}$$

$$1 = \sum_{n=0}^{\infty} P_i(n). \tag{7}$$

The sum of higher order state probabilities for the revised model cannot be represented by a geometric series, so closed-form expressions for computing performance measures do not exist. However, the death rate for the process must eventually exceed the birth rate as n increases, at which point the state probabilities decrease faster than geometrically. Therefore, we can truncate the state space to an upper bound N that limits excluded probability to a value less than an arbitrary parameter ε (e.g., 10^{-5}). Letting $\rho = \lambda_i/\{s_i/h_i + (N-s_i)/b\}$, it is clear that

$$\sum_{n=N+1}^{\infty} P_i(n) < P_i(N)\sum_{n=1}^{\infty}\rho^n = \frac{P_i(N)\rho}{1-\rho} \tag{8}$$

for $\rho < 1$. Hence, the following algorithm determines a truncation state N and state probabilities $P_i(n), n \in \{0,\ldots,N\}$, such that $\sum_{n=0}^{N} P_i(n) = 1$ and redistributed probability is less than ε:

```
$P_i(s_i) = 1$
$C = 1$
for $n = s_i - 1$ down to 0
    $P_i(n) = (n+1)P_i(n+1)/(\lambda_i h_i)$
    $C = C + P_i(n)$
next $n$
$N = s_i$
$\rho = \lambda_i h_i / s_i$
do while $\rho P_i(N) > (1-\rho)C\varepsilon$
    $N = N+1$
    $\rho = \lambda_i / \{(s_i/h_i) + (N - s_i)/b\}$
    $P_i(N) = \rho P_i(N-1)$
    $C = C + P_i(\mathrm{N})$
loop
for $n = 0$ to $N$
    $P_i(n) = P_i(n)/C$
next $n$.
```

Note that the algorithm initiates pre-normalized probability computation at a state with relatively high expected probability (s_i), since numerical problems can occur when anchoring on a low probability state such as 0 (see Smith [29]). With state probabilities determined, expected queue length for interval i can be calculated as

$$Q_i = \sum_{n = s_i + 1}^{N} (n - s_i) P_i(n) \tag{9}$$

and the fraction of abandoning customers can be derived as the ratio of the average abandonment rate Q_i/b over the customer arrival rate; that is,

$$B_i = Q_i/(b\lambda_i). \tag{10}$$

To compute the interval service level L_i, we adapt a compact recursive implementation of the foundational result derived by Riordan [23] (see Avramidis et al. [30], p. 487):

```
$D = P_i(s_i)$
$F = 1$
$G = 1$
for $n = s_i + 1$ to $N$
    $G = G\{(s_i b/h_i) + n - s_i - 1\}\{1 - \exp(-t/b)\}(n - s_i)$
    $F = F + G$
    $D = D + FP_i(n)$
next $n$
$L_i = \{1 - D\exp(-s_i t/h_i)\}(1 - B_i)$.
```

The last step of the algorithm enforces an optional rule (employed at Qwest) that only non-abandoning customers can contribute positively to service level (that is, service level for any interval i can be at most $1-B_i$). For noninteger staffing levels (a consequence of applying an availability factor to scheduled staff), values of Q_i, B_i, and L_i can be obtained by interpolating between corresponding results for $\lfloor s_i \rfloor$ and $\lceil s_i \rceil$. When determining an interval staffing requirement r_i, we can calculate an initial service level based on $s_i = [\lambda_i h_i]$. We then increment or decrement s_i as needed until the target service level is bracketed by the last two results, and interpolate linearly between them to obtain r_i. Since the expected available staff on an interval need not be integer valued, similar treatment of required staff is analytically appropriate. Additional justification for non-integrality of r_i follows from computational reliance on expected values for the contributing parameters, which will differ from the actual values observed in any particular implementation.

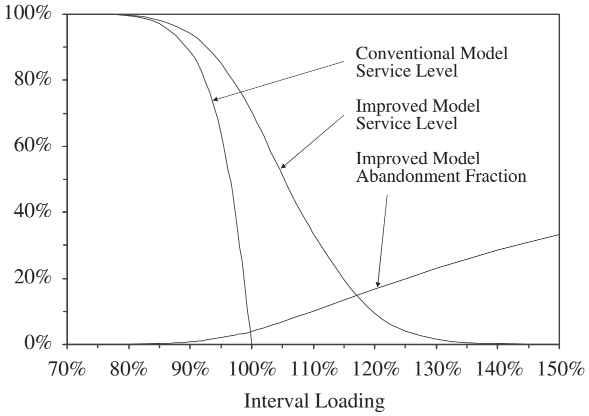


Fig. 3. Comparison of queueing model results.

Fig. 3 compares typical performance results for the Markovian queueing models with and without customer abandonment. For this illustration, $t = 20$ s, $h_i = 300$ s, $b = 300$ s (where applicable), $s_i = 100$, and λ_i is parameterized to produce a range of loading levels $\lambda_i h_i / s_i$ between 70% and 150%. The comparatively high sensitivity of the conventional model is noteworthy, with service level dropping from nearly 100% at 80% loading to 0 at 100% loading. In contrast, the improved model indicates a service level of about 70% at 100% loading with about 4% of customers abandoning. At 150% loading, the service level has decayed to nearly zero and, as we would expect, one-third of the arriving customers abandon the queue. Even though abandoning customers cannot themselves be served, they significantly improve service level performance by reducing system congestion precisely under those conditions when the reduction is most beneficial.

All of the staffing computations described in this section can be easily implemented in a spreadsheet environment, with VBA modules employed to execute the algorithmic procedures. For typical input parameters and $\varepsilon = 10^{-5}$, the truncation state N is less than twice the magnitude of r_i for a representative interval i. Hence, the computation time required to obtain performance results for all intervals is typically less than 1 s.

3. Schedule optimization

In any given week, repair service agents at Qwest must be scheduled in shift groupings called "tours" that span multiple intervals. Each tour is characterized by its combination of start time, workday schedule, and shift pattern (standard or split). The standard shift consists of 8.5 consecutive hours (8 h of work with 30 min for lunch), whereas the split shift consists of 12 consecutive hours (4 h of work, 4 h off, and another 4 h of work). The split tours are included in the eligible tour set because standard shifts alone tend to provide an inefficient set of basis functions for accommodating the call volume profile. Not surprisingly, agents generally prefer standard tours, so our model includes the capability to limit the portion of agents p who are assigned to split tours. Eligible start times align with a one-day interval lattice so that specified hours of operation are maintained. For a call center open between 6 AM (interval 13) and 11 PM (interval 47), the eligible start times T_j would include $\{13,14,\ldots,30\}$ for a standard tour j, and $\{13,14,\ldots,23\}$ for a split tour. A coverage parameter c_{ij} indicates which intervals i are covered by tour type j when artificially assuming a start time of midnight (interval 1). This parameter is generally binary, though we represent the lunch period in each standard tour with a 90-min notch $(\ldots,1.00,1.00,0.75,0.50,0.75,1.00,1,00,\ldots)$ rather than a 30-min slot $(\ldots,1.00,1.00,0.00,$

1.00,1,00, . . .) to acknowledge the likely tactical staggering of lunch periods for agents scheduled for the same shift. Lunches could be staggered explicitly by introducing more tour types, but the resulting schedules would be far more complex and could create agent expectations that inhibit tactical flexibility in responding to unpredicted fluctuations in workload (for example, we may want to delay some lunches until correction of a major outage that affects service in a wide geographic area).

Diversity of available tours ensures that each agent works five days in each week, Saturdays and Sundays are adequately staffed, and higher volume days employ more agents. In any feasible schedule, some agents may be assigned to simple Monday through Friday tours (from the set denoted R_5 for standard tours or S_5 for split tours). Other agents will be assigned a Saturday or Sunday tour (from the set R_1 for standard or S_1 for split), along with a weekday tour with a nonscheduled (NS) day (from the set R_4 for standard or S_4 for split). Each agent's weekday start time will not vary within the week. However, an agent assigned to an NS day tour may have a completely different tour type and start time on the assigned Saturday or Sunday.

Our modeling objective is to generate a tour distribution such that a staffing level of at least r_i is achieved on each interval, while minimizing the sum of squared normalized deviations between available staff and required staff. This objective recognizes the "diminishing returns" property of performance improvement as staffing increases, and therefore distributes staff surpluses as evenly as possible throughout the week. The ancillary restriction on the frequency of split tours p must also be enforced, along with permissibility of weekend split tours as controlled by binary parameter q (1=yes, 0=no). To capture these requirements, we formulate a quadratic program in which the decision variable x_{jk} is the number of agents assigned to tour j with start time $k \in T_j$. Letting auxiliary variable y_i be the staffing level for interval i, we write the formulation

$$\text{Minimize} \quad \sum_{i \in I} \left(\frac{y_i - r_i}{r_i} \right)^2 \tag{11}$$

$$\text{Subject to} \quad y_i = a_i \sum_{j \in R \cup S} \sum_{k \in T_j} c_{i-k+1,j} x_{jk}, \quad i \in I \tag{12}$$

$$y_i \geq r_i, \; i \in I \tag{13}$$

$$\sum_{j \in R_4 \cup S_4} \sum_{k \in T_j} x_{jk} = \sum_{j \in R_1 \cup S_1} \sum_{k \in T_j} x_{jk} \tag{14}$$

$$\sum_{j \in S_5 \cup S_4} \sum_{k \in T_j} x_{jk} \leq p \sum_{j \in R_5 \cup R_4 \cup S_5 \cup S_4} \sum_{k \in T_j} x_{jk} \tag{15}$$

$$\sum_{j \in S_1} \sum_{k \in T_j} x_{jk} \leq pq \sum_{j \in R_1 \cup S_1} \sum_{k \in T_j} x_{jk} \tag{16}$$

$$x_{jk} \geq 0, \quad j \in R \cup S, \;\; k \in T_j \tag{17}$$

where $R = R_5 \cup R_4 \cup R_1$ and $S = S_5 \cup S_4 \cup S_1$. By employing standard optimization software, we can solve the quadratic program in a fraction of a second on a personal computer. The optimal distribution of tour types and start times is then given by

$$z_{jk} = x_{jk} \left(\sum_{l \in R_5 \cup R_4 \cup S_5 \cup S_4} \sum_{m \in T_l} x_{lm} \right)^{-1}, \quad j \in R \cup S, k \in T_j. \tag{18}$$

To find an optimal schedule for a specified number of agents M, we must first scale the optimal distribution given by Eq. (18) and then convert the scaled result to an integer schedule. Enforcing integer solutions within the formulation is impractical because of the employment of a nonlinear objective function. Naive rounding is not appropriate because there is no guarantee that

$$\sum_{j \in R_5 \cup R_4 \cup S_5 \cup S_4} \sum_{k \in T_j} [M z_{jk}] = M. \tag{19}$$

Similarly, equality of NS and weekend tour quantities is not assured; that is, we cannot be certain that

$$\sum_{j \in R_4 \cup S_4} \sum_{k \in T_j} [M z_{jk}] = \sum_{j \in R_1 \cup S_1} \sum_{k \in T_j} [M z_{jk}], \tag{20}$$

which is a requirement for a feasible integer schedule. Fortunately, we can perform successive one-dimensional searches to find upward rounding thresholds such that each of these conditions is satisfied.

Fig. 4 illustrates the rounding process for a typical Qwest scheduling problem with M=360 agents. For this realization, upward rounding of all $M z_{jk}, j \in R_5 \cup R_4 \cup S_5 \cup S_4, k \in T_j$, (threshold=0.0) would schedule tours for 421 agents, whereas downward rounding (threshold=1.0) would schedule tours for 316 agents. The relationship between the upward rounding threshold and the resulting number of scheduled tours is obviously monotonic, so a simple one-dimensional search algorithm such as interval bisection can be employed to quickly obtain a rounding threshold that satisfies Eq. (19). Based on this threshold, the number of scheduled NS day tours is 175. We then apply a similar process to produce a rounding threshold for $M z_{jk}, j \in R_1 \cup S_1, k \in T_j$, that schedules 175 weekend tours and therefore satisfies Eq. (20).

Letting $x'_{jk}, j \in R \cup S, k \in T_j$, be the integer schedule produced by the rounding process, the resulting staffing level for each interval i is

$$y'_i = a_i \sum_{j \in R \cup S} \sum_{k \in T_j} c_{i-k+1,j} x'_{jk}. \tag{21}$$

Due to imperfect correlation between interval work volumes and staffing levels in any implemented schedule, the variability of a_i necessitates slight scaling of y'_i to ensure that the sum of available staff across intervals is consistent with $A \times M$. Letting K be the number of intervals each agent covers during a week (typically 80),

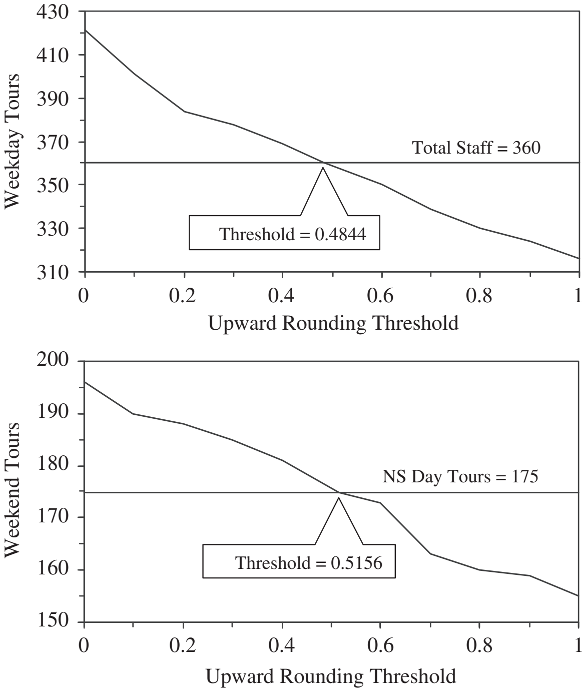


Fig. 4. Determination of rounding thresholds.

corrected available staff for each interval i can be computed as

$$s_i = AMKy'_i/\sum_{j\in I} y'_j. \tag{22}$$

Interval performance measures Q_i, B_i, and L_i are then determined through the methods described in the previous section. In addition, available staff utilization is calculated as

$$U_i = \lambda_i h_i(1-B_i)/s_i. \tag{23}$$

Composite performance measures for the entire week can then be derived as

$$U = \sum_{i\in I} \lambda_i h_i(1-B_i)/(AMK), \tag{24}$$

$$Q = \sum_{i\in I} Q_i/|I|, \tag{25}$$

$$B = \sum_{i\in I} f_i^v B_i, \tag{26}$$

$$L = \sum_{i\in I} f_i^v L_i. \tag{27}$$

The efficiency of the schedule is $\sum_{i\in I} r_i/\sum_{i\in I} s_i = \sum_{i\in I} r_i/(AMK)$, since a notional schedule where $s_i = r_i, i \in I$, will be perfectly efficient.

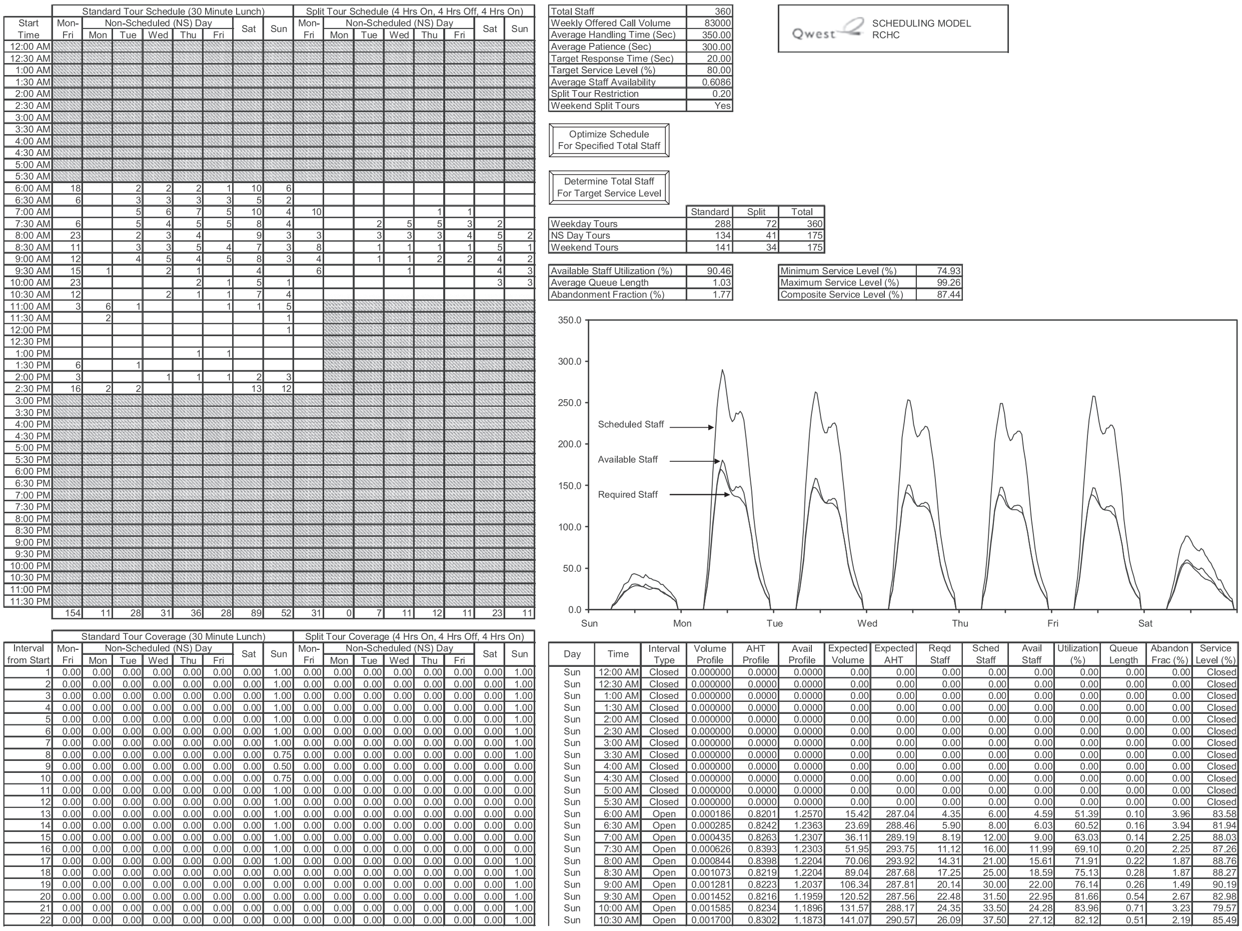
It should be emphasized that, when the total number of agents M is prespecified, the realized composite service level may vary substantially from the desired target (the model maximizes the service level given the available resources). While our normal operational objective is to optimally schedule a fixed number of agents, we can also employ the model to determine the number of agents required to achieve a target service level within the strategic planning context described earlier. Various approaches for addressing this problem have been offered by several authors including Buffa et al. [7] and Koole and van der Sluis [31], but the parsimony and high solution speed of our model suggest a very simple algorithm. We begin by determining the total number of agents required when scheduling efficiency is artificially assumed to be 100%; that is, let

$$M = \left\lceil \sum_{i\in I} r_i/(AK) \right\rceil. \tag{28}$$

We then increment M and re-optimize the schedule until L exceeds the target service level. With this approach, the required number of agents and associated optimal schedule can be determined within a few seconds of spreadsheet computation time.

4. Implementation and results

The complete method for determining an optimal schedule and predicting performance has been implemented in a Microsoft Excel spreadsheet environment. Fig. 5 displays key components of the model, including a schedule matrix, global input parameters, a performance summary, a coverage parameter matrix, and interval performance results. The offered weekly call volume, which is one of the global inputs, is generated from a forecasting model that annualizes the volume (based on a seasonality profile), employs exponential smoothing on annualized volume, and then reapplies



	Standard Tour Schedule (30 Minute Lunch)								Split Tour Schedule (4 Hrs On, 4 Hrs Off, 4 Hrs On)							
Start Time	Mon-Fri	Non-Scheduled (NS) Day					Sat	Sun	Mon-Fri	Non-Scheduled (NS) Day					Sat	Sun
		Mon	Tue	Wed	Thu	Fri				Mon	Tue	Wed	Thu	Fri		
12:00 AM																
12:30 AM																
1:00 AM																
1:30 AM																
2:00 AM																
2:30 AM																
3:00 AM																
3:30 AM																
4:00 AM																
4:30 AM																
5:00 AM																
5:30 AM																
6:00 AM	18		2	2	2	1	10	6								
6:30 AM	6		3	3	3	3	5	2								
7:00 AM			5	6	7	5	10	4	10				1	1		
7:30 AM	6		5	4	5	5	8	4			2	5	5	3	2	
8:00 AM	23		2	3	4		9	3	3		3	3	3	4	5	2
8:30 AM	11		3	3	5	4	7	3	8		1	1	1	1	5	1
9:00 AM	12		4	5	4	5	8	3	4		1	1	2	2	4	2
9:30 AM	15	1		2	1		4		6			1			4	3
10:00 AM	23				2	1	5	1							3	3
10:30 AM	12			2	1	1	7	4								
11:00 AM	3	6	1			1	1	5								
11:30 AM		2						1								
12:00 PM								1								
12:30 PM																
1:00 PM					1	1										
1:30 PM	6		1													
2:00 PM	3			1	1	1	2	3								
2:30 PM	16	2	2				13	12								
3:00 PM																
3:30 PM																
4:00 PM																
4:30 PM																
5:00 PM																
5:30 PM																
6:00 PM																
6:30 PM																
7:00 PM																
7:30 PM																
8:00 PM																
8:30 PM																
9:00 PM																
9:30 PM																
10:00 PM																
10:30 PM																
11:00 PM																
11:30 PM																
	154	11	28	31	36	28	89	52	31	0	7	11	12	11	23	11

Total Staff	360
Weekly Offered Call Volume	83000
Average Handling Time (Sec)	350.00
Average Patience (Sec)	300.00
Target Response Time (Sec)	20.00
Target Service Level (%)	80.00
Average Staff Availability	0.6086
Split Tour Restriction	0.20
Weekend Split Tours	Yes

Qwest SCHEDULING MODEL
RCHC

	Standard	Split	Total
Weekday Tours	288	72	360
NS Day Tours	134	41	175
Weekend Tours	141	34	175

Available Staff Utilization (%)	90.46
Average Queue Length	1.03
Abandonment Fraction (%)	1.77

Minimum Service Level (%)	74.93
Maximum Service Level (%)	99.26
Composite Service Level (%)	87.44

	Standard Tour Coverage (30 Minute Lunch)								Split Tour Coverage (4 Hrs On, 4 Hrs Off, 4 Hrs On)							
Interval from Start	Mon-Fri	Non-Scheduled (NS) Day					Sat	Sun	Mon-Fri	Non-Scheduled (NS) Day					Sat	Sun
		Mon	Tue	Wed	Thu	Fri				Mon	Tue	Wed	Thu	Fri		
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Day	Time	Interval Type	Volume Profile	AHT Profile	Avail Profile	Expected Volume	Expected AHT	Reqd Staff	Sched Staff	Avail Staff	Utilization (%)	Queue Length	Abandon Frac (%)	Service Level (%)
Sun	12:00 AM	Closed	0.000000	0.0000	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Closed
Sun	12:30 AM	Closed	0.000000	0.0000	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Closed
Sun	1:00 AM	Closed	0.000000	0.0000	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Closed
Sun	1:30 AM	Closed	0.000000	0.0000	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Closed
Sun	2:00 AM	Closed	0.000000	0.0000	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Closed
Sun	2:30 AM	Closed	0.000000	0.0000	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Closed
Sun	3:00 AM	Closed	0.000000	0.0000	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Closed
Sun	3:30 AM	Closed	0.000000	0.0000	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Closed
Sun	4:00 AM	Closed	0.000000	0.0000	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Closed
Sun	4:30 AM	Closed	0.000000	0.0000	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Closed
Sun	5:00 AM	Closed	0.000000	0.0000	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Closed
Sun	5:30 AM	Closed	0.000000	0.0000	0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Closed
Sun	6:00 AM	Open	0.000186	0.8201	1.2570	15.42	287.04	4.35	6.00	4.59	51.39	0.10	3.96	83.58
Sun	6:30 AM	Open	0.000285	0.8242	1.2363	23.69	288.46	5.90	8.00	6.03	60.52	0.16	3.94	81.94
Sun	7:00 AM	Open	0.000435	0.8263	1.2307	36.11	289.19	8.19	12.00	9.00	63.03	0.14	2.25	88.03
Sun	7:30 AM	Open	0.000626	0.8393	1.2303	51.95	293.75	11.12	16.00	11.99	69.10	0.20	2.25	87.26
Sun	8:00 AM	Open	0.000844	0.8398	1.2204	70.06	293.92	14.31	21.00	15.61	71.91	0.22	1.87	88.76
Sun	8:30 AM	Open	0.001073	0.8219	1.2204	89.04	287.68	17.25	25.00	18.59	75.13	0.28	1.87	88.27
Sun	9:00 AM	Open	0.001281	0.8223	1.2037	106.34	287.81	20.14	30.00	22.00	76.14	0.26	1.49	90.19
Sun	9:30 AM	Open	0.001452	0.8216	1.1959	120.52	287.56	22.48	31.50	22.95	81.66	0.54	2.67	82.98
Sun	10:00 AM	Open	0.001585	0.8234	1.1896	131.57	288.17	24.35	33.50	24.28	83.96	0.71	3.23	79.57
Sun	10:30 AM	Open	0.001700	0.8302	1.1873	141.07	290.57	26.09	37.50	27.12	82.12	0.51	2.19	85.49

Fig. 5. Spreadsheet implementation of the scheduling model.

the seasonality effect. Tours can be entered manually and the model will check for feasibility, alerting the user to discrepancies through warning messages. An optimal schedule can be generated by a single mouse click on the button beneath the global parameters. A second button initiates the iterative process for determining the number of agents required to achieve the input target service level. Any optimal schedule can be modified manually, and the effect on predicted performance can be immediately observed.

Error checking (type and range) is performed on all data inputs. The split tour restriction input is selected from a drop-down list, along with the binary input for permissibility of weekend split tours. Optimal tour distributions are stored for each combination of restrictions, which are created using baseline values for the global input parameters. We employ this approach because, while realistic variations in these values affect the scale of the required staff curve, they do not appreciably affect its shape. Operationally, the stored solutions are updated whenever new profiles are entered for call volume, average handling time, and agent availability (typically, profiles are updated at least every six months for stable products and associated centers). The resulting simple, portable model has the added advantage of producing schedules that do not exhibit radically different structures from one scheduling period to the next. For each period, tours from the optimal schedule are assigned to individual agents in seniority order based on submitted preference rankings for all tours.

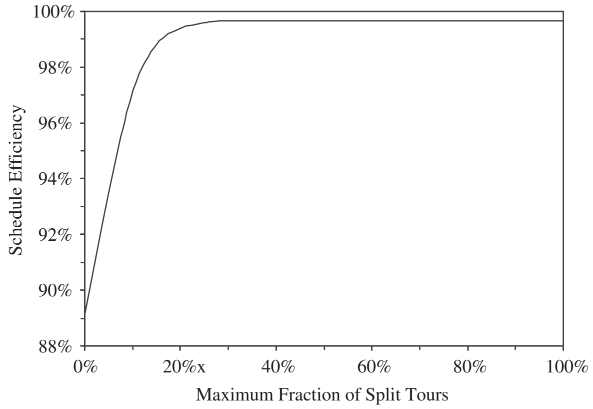


Fig. 6. Value of split tours.

The Qwest call center represented in Fig. 5 is open from 6 AM to 11 PM daily (overnight calls are routed to a single consolidated center that handles multiple products). For the given input parameters, 360 scheduled agents ensure that 87.44% of offered calls are answered within 20 s. For interval performance, the model predicts minimum and maximum service levels of 74.93% and 99.26%, respectively. The overall customer abandonment fraction is 1.77%, and the maximum abandonment fraction for any interval is 5.03%. These abandonment predictions align well with empirical experience.

The above results are based on a split tour restriction of 20%. As indicated by Fig. 6, some split tours must be employed to produce reasonably efficient schedules. With split tours completely prohibited, a scheduling efficiency of 89.16% is realized and 389 agents are needed to achieve the target service level of 80%. With split tours completely unrestricted, efficiency increases to 99.67% and only 348 agents are required. For the unrestricted case, the model schedules 90 weekday split tours (26%) and 63 weekend split tours (38%). However, as the figure illustrates, nearly all of the efficiency improvement can be achieved with no more than 20% split tours. In practice, the model is extremely valuable in providing this type of insight.

We also compared our optimization method with the traditional approach, which is to employ an integer programming model and minimize the total number of scheduled agents while enforcing staff requirements on all intervals. Unless the schedule is perfectly efficient, some intervals will be overstaffed and the composite service level will exceed the target. The interval service level requirement can be iteratively modified to converge on a composite target, but a new integer program must be solved at each iteration. For the operationally normal case where the total number of agents is specified, the agent population could be fixed by a constraint and a different linear objective function could be employed. For example, we could minimize the maximum normalized surplus encountered on any interval. From our integer programming experiments, we observe that the most effective approach is to retain the objective of minimizing total staff, and iteratively adjust the interval service level requirement (resolving the model) until the optimal objective value is equal to the total

Table 1
Comparison of optimization methods.

	IP minimum service level (%)				Split tour limit (%)
	50	60	70	80	
Total staff	391	402	415	430	0
IP CPU time (s)/Gap (%)	1/0.00	1/0.00	1/0.00	1/0.00	
IP composite service level (%)	75.57	82.10	86.96	91.80	
QP composite service level (%)	80.60	85.05	89.57	93.59	
QP minimum service level (%)	44.18	49.94	66.37	74.99	
Total staff	353	363	374	388	10
IP CPU time (s)/Gap (%)	1000/0.34	1/0.00	1/0.00	1000/0.39	
IP composite service level (%)	75.60	79.94	86.36	91.49	
QP composite service level (%)	78.30	83.75	88.75	93.39	
QP minimum service level (%)	47.00	46.98	62.24	74.94	
Total staff	328	338	348	360	20
IP CPU time (s)/Gap (%)	258/0.00	1000/0.39	1000/0.37	1000/0.29	
IP composite service level (%)	61.56	70.94	78.48	86.21	
QP composite service level (%)	63.34	72.49	79.97	87.44	
QP minimum service level (%)	42.57	51.67	62.78	74.93	
Total staff	327	336	346	358	100
IP CPU time (s)/Gap (%)	1000/0.45	1000/0.42	1000/0.46	1000/0.49	
IP composite service level (%)	61.11	68.27	77.35	85.65	
QP composite service level (%)	62.87	71.13	79.12	86.62	
QP minimum service level (%)	36.93	51.36	62.10	74.02	

number of agents to be scheduled. Unfortunately, unlike our quadratic programming objective, none of the implementable integer programming objectives captures the nonlinear "diminishing returns" property of queueing system performance.

Table 1 compares the service level performance of the quadratic programming approach (QP) against a conventional integer program (IP), parameterizing on the minimum interval service level (which determines $r_i, i \in I$, for both methods) and the allowable fraction of split tours. In this comparison, the total staff quantity is determined by the integer programming optimization, and this number is then employed as M in the competing quadratic programming solution. All mathematical programs are solved using CPLEX 11.2 software on a personal computer with a 2.26 GHz processor. The integer programs often require many hours of CPU time to prove optimality (sometimes running out of memory), so run times are restricted to 1000 s with the percent gap between upper and lower bounds documented in the table (the gap always closes to less than 1% within about 1 s, but we extend the run time to strengthen the quality of the investigative comparison). In every case shown, any suboptimality due to rounding is dominated by the modeling advantage of the quadratic objective function in maximizing the composite service level. The differences range from 0.96% to 5.03%, with the rounded quadratic result averaging 2.22% better than the integer programming result for the same number of scheduled agents. The differentiation is generally more pronounced when split tours are more severely restricted, since the restrictions lead to less efficient schedules and, consequently, larger staffing surpluses to be distributed (the integer programming method distributes surpluses arbitrarily, without regard for diminishing returns). While the table indicates lower minimum service levels for the quadratic programming approach, the average decrease of 7.63% from the integer programming baseline is not excessive. Uniformity of service levels across intervals is secondarily desirable, but the composite service level is the primary performance metric.

It should be noted that, while the interval staffing constraint imposed by Eq. (13) is required for the integer program, it is optional for the quadratic programming approach. In preliminary studies, we experimented with eliminating this constraint (and consequently minimizing the sum of both squared normalized surpluses and squared normalized shortages). This modification yields slight improvement in composite service levels. However, very low service levels can occur for some intervals under typical scenarios, resulting in excessive nonuniformity of service and weakened validity of the SIPP assumption. For the example application shown in Fig. 5, removing the constraint marginally increases the composite service level from 87.44% to 87.84%, but markedly decreases the minimum service level from 74.93% to 52.02%. By retaining the constraint, we achieve more balanced schedules by imposing a much stronger impediment to shortages than to surpluses.

5. Concluding remarks

We have integrated queueing theory, quadratic programming, and a variable-threshold rounding algorithm to develop a practical, spreadsheet-based model for call center scheduling. The model has been successfully implemented in several repair service centers at Qwest Communications, resulting in substantial cost reductions and near elimination of service level target misses. With the new approach, we observe a 15–20% reduction in personnel requirements from those produced by previous scheduling methods. For some centers, the model has been expanded to accommodate additional tour types and varying operational practices. Hence, our experience suggests that the approach is quite flexible and could be applied to myriad call center environments including other repair services, product delivery, public services, and retail sales.

References

[1] Mandelbaum A. Quality and efficiency driven queues (with a focus on call/contact centers). Euro Working Group on Stochastic Modeling, Koc University; June 23–25 2008.

[2] Gans N, Koole G, Mandelbaum A. Telephone call centers: tutorial, review, and research prospects. Manufacturing and Service Operations Management 2003;5(1):79–141.

[3] Mandelbaum A. Call centers (centres): research bibliography with abstracts, Version 7, 〈http://ie.technion.ac.il/serveng〉; 2006.

[4] Andrews BH, Parsons HL, Bean LL. Chooses a telephone agent scheduling system. Interfaces 1989;19(6):1–9.

[5] Linder RW. The development of manpower and facilities planning methods for airline telephone reservation offices. Operational Research Quarterly 1976;20(1):3–21.

[6] Harris CM, Hoffman KL, Saunders PB. Modeling the IRS telephone taxpayer information system. Operations Research 1987;35(4):504–23.

[7] Buffa ES, Cosgrove MJ, Luce BJ. An integrated work shift scheduling system. Decision Sciences 1976;7(4):620–30.

[8] Church JG. Sure Staf: a computerized staff scheduling system for telephone business offices. Management Science 1973;20(4):708–20.

[9] Segal M. The operator-scheduling problem: a network-flow approach. Operations Research 1974;22(4):808–23.

[10] Thompson GM. Improved implicit optimal modeling of the labor shift scheduling problem. Management Science 1995;41(4):595–607.

[11] Sze DY. A queueing model for telephone operator scheduling. Operations Research 1984;32(2):229–49.

[12] Muhlemann AP. A simulation study of the operations of a telephone bureau. Omega 1981;9(6):633–7.

[13] Gans N, Zhou Y. A call routing problem with service-level constraints. Operations Research 2003;51(2):255–71.

[14] Ingolfsson A, Haque A, Umnikov A. Accounting for time-varying queueing effects in workforce scheduling. European Journal of Operational Research 2002;139(3):585–97.

[15] Henderson WB, Berry WL. Heuristic methods for telephone operator shift scheduling. Management Science 1976;22(12):1372–80.

[16] Brigandi AJ, Dargon DR, Sheehan MJ, Spencer T. AT&T's call processing simulator (CAPS) operational design for inbound call centers. Interfaces 1994;24(1):6–28.

[17] Green LV, Kolesar PJ, Soares J. Improving the SIPP approach for staffing service systems that have cyclic demands. Operations Research 2001;49(4):549–64.

[18] Green LV, Kolesar PJ, Whitt W. Coping With time-varying demand when setting staffing requirements for a service system. Production and Operations Management 2007;16(1):13–39.

[19] Cooper RB. Introduction to queueing theory. 2nd ed. New York: Elsevier; 1981.

[20] Dietz DC, Vaver JG. Synergistic modeling of call center operations. Journal of Applied Mathematics and Decision Sciences 2006:1–13. (Article ID 53928).

[21] Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, et al. Statistical analysis of a telephone call center: a queueing science perspective. Journal of the American Statistical Association 2005;100(469):36–50.

[22] Palm C. Etude Des Delais D'Attente. Ericsson Technics 1937;5(1):37–56.

[23] Riordan J. Stochastic service systems. New York: Wiley; 1962.

[24] Garnett O, Mandelbaum A, Reiman M. Designing a call center with impatient customers. Manufacturing and Service Operations Management 2002;4(3):208–27.

[25] Stolletz R. Performance analysis and optimization of inbound call centers. Berlin: Springer; 2003.

[26] Feldman Z, Mandelbaum A, Massey WA, Whitt W. Staffing of time-varying queues to achieve time-stable performance. Management Science 2008;54(2):324–38.

[27] Whitt W. What you should know about queueing models to set staffing requirements in service systems. Naval Research Logistics 2007;54(5):476–84.

[28] Mandelbaum A. Zeltyn S. Service engineering in action: the Palm/Erlang-A queue, with applications to call centers. In: Spath D., Fahnrich K, editors, Advances in services innovations. Berlin: Springer; 2007. p. 17–45.

[29] Smith DK. Calculation of steady-state probabilities of M/M queues: further approaches. Journal of Applied Mathematics and Decision Sciences 2002;6(1): 43–50.

[30] Avramidis AN, Chan W, L'Ecuyer P. Staffing multi-skill call centers via search methods and a performance approximation. IIE Transactions 2009;41(6):483–97.

[31] Koole G, van der Sluis E. Optimal shift scheduling with a global service level constraint. IIE Transactions 2003;35(11):1049–55.